

## Answers to Reliability-Validity Knowledge Check Questions

1. Weighing yourself on a scale 3 times and getting the following readings: 150 lbs., 157 lbs., 153 lbs.

**This example primarily illustrates low reliability: the scale is yielding inconsistent output (a 7 pound range) simply by getting on and off the scale three times. Measures with low reliability always have low validity as well. Although the construct of “weight” has validity, this scale could not provide a valid measure of weight because it doesn’t even yield consistent measurements in the first place.**

2. Administering a job skills test to 100 job applicants, hiring the 50 best scorers, and then finding out that even among these 50 new employees, those who scored higher on the job skills test tend to perform better on the job.

**Criterion validity focuses on how well a measure predicts some future or current criterion (usually, behavior). In this example, the job skills test has good criterion validity because it predicts future job performance: Higher scores on the test were predictive of better on-the-job performance. (Secondarily, if supported by other findings, this good criterion validity will help to establish the construct validity of this job skills test as well.) Note that in terms of establishing criterion validity, it would have been ideal if the company had hired all 100 job applicants – even those who scored poorly on the skills test – and then examined how well the test predicted actual job performance. In the real world, however, companies are not likely to do this.**

3. Students who score in the top 10% on the ACT (a college aptitude test) tend to score in about the same percentile on the SAT (a different college aptitude test).

**In this example, "college aptitude" is an underlying construct (concept) that we are interested in, and we have two measures of it: ACT scores and SAT scores. Most directly, this example illustrates high convergent validity, because the ACT and SAT are both supposed to be measuring college aptitude and therefore they should yield similar results. Higher scores on one test should correlate substantially with higher scores on the other test. This high convergent validity helps to establish the construct validity of these aptitude tests.**

4. After many administrations, researchers administering a polygraph test begin to worry that the machine is actually measuring anxiety and not dishonest responses.  
**This illustrates a concern about potentially low construct validity, because the concern is that the instrument (the polygraph) does not appear to be measuring the desired construct (dishonesty), but is instead measuring a different theoretical construct (anxiety).**
  
5. A personality test that helps to predict the development of schizophrenia consists entirely of items such as “What is your favorite color?” and “Are red apples better than green apples?”  
**In this hypothetical example, the personality test has low face validity, because the items on the test seem to be unrelated to the construct (schizophrenia) being measured. What on Earth do favorite colors and “red versus green apples” have to do with schizophrenia? But even though the items might look silly or irrelevant, the more important issue (in terms of developing psychological tests) is that the test has high criterion validity: based on the information provided in the example, the test helps to predict the development of schizophrenia.**
  
6. Individuals that score high on a questionnaire measuring racism on Tuesday morning are likely to score high on the same scale one week later.  
**This illustrates high reliability, because multiple administrations of the questionnaire are yielding similar results. Note that the questionnaire’s high reliability does not indicate anything about its validity. The questionnaire might have low or high validity – we need more information to determine this.**
  
7. Two researchers use a newly developed observational coding system to record how newlywed partners interact with one another. The researchers' goal is to predict which couples will be divorced within 4 years. Results show that the "behavioral profiles" established by the coding system correctly predicted divorce 90% of the time.  
**This coding system has high criterion validity, because it is highly successful in predicting future divorce. (Ultimately, if supported by other types of evidence, high criterion validity will help to establish the general construct validity of this coding scale. In other words, it will help to establish that this coding system really is measuring marital conflict and unhappiness).**

8. Research consistently shows that scores on Dr. Smith's "Selfishness Test" are highly positively correlated with scores on other selfishness tests and, as hypothesized, are moderately correlated with scores on tests that measure "egocentrism."  
**Most directly, these findings indicate that Dr. Smith's selfishness scale has high convergent validity. First, it is highly correlated with other measures of the same construct (i.e., the other psychological tests that measure selfishness). Second, let's assume that based on existing psychological theories, egocentrism and selfishness are not considered identical traits, but they are constructs that should be somewhat related to one another. In this case, the fact that Dr. Smith's selfishness scale correlates moderately with psychological tests of egocentrism shows, once again, good convergent validity. This overall pattern of convergent validity supports the overall construct validity of Dr. Smith's selfishness measure.**
9. Suppose shyness and extraversion should be *negatively* correlated. We find that people with higher scores on a new shyness test (indicating they are more shy) also score higher on a previously validated extraversion test (indicating they are more socially outgoing). **The new shyness test has low (poor) convergent validity, which in turn suggests *poor construct validity*. In other words, based on psychological theory, we would expect that people who have higher scores on the new shyness test should generally have lower scores on extraversion: they should be less socially outgoing. Thus, the scores from the two tests should converge – should be related -- but in an opposite direction (remember, a negative and positive correlation of any particular value are just as strong – show the same degree of relation or convergence; it's only the direction that differs). Thus, we would expect a negative correlation between shyness and extraversion, but this isn't what happened: the correlation was positive. If other studies yield similar findings, this suggests that our new test has poor construct validity and is not measuring shyness.**
10. Students in Professor Jones' Geography 215 class are assigned to read Chapters 1, 2, 3, and 4 for the first exam. All the chapters are similar in length and amount of material. In lecture Professor Jones conducts 3 lectures on the topics in each chapter. Students are told to study all chapters and lecture notes for their first exam. On the first exam, however, 90% of the exam questions are based on the material in Chapter 3 and Chapter 4, and only 10% of the questions are based on material in Chapter 1 and Chapter 2. **Most directly this example illustrates that Professor Jones' exam has low content validity. The sample of questions contained in the exam poorly represents the domain of material that students were asked to read about and which they learned about in lecture. Roughly 25% of the material covered in class and the text focused on concepts related to Chapter 1, and another 25% was related to Chapter 2. Yet only 10% of the exam questions focused on topics from these two chapters combined. Secondly, the poor content validity likely will cause many students to feel that this was not a "fair exam." If so, then poor content validity would lead to *poor face validity*.**